Instruction-tuning Aligns LLMs to the Human Brain

Khai Loong Aw, Syrielle Montariol*, **Badr AlKhamissi***, Martin Schrimpf+, Antoine Bosselut+

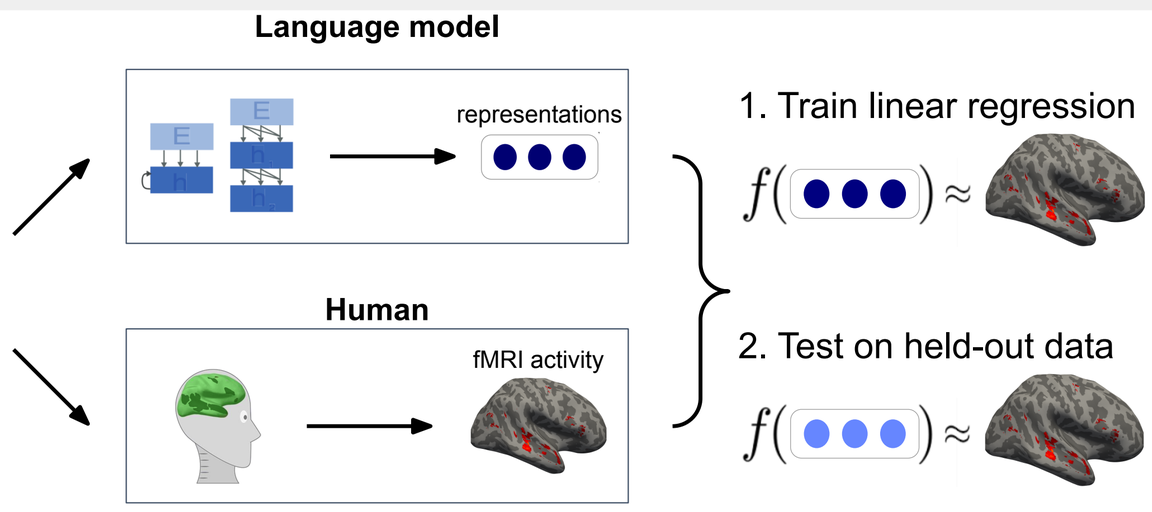
*Equal contribution, +Equal supervision / senior authors

Preliminaries

Brain alignment: Evaluate similarity of LLM internal representations to human brain activity

Language stimuli

Pereira2018	"Beekeeping encourages the conservation of local habitats. It is in every beekeeper's ..."
Blank2014	"If you were to journey to the North of England, you would come to a valley that ..."
Wehbe2014	"Harry had never believed he would meet a boy he hated more than Dudley, but ..."



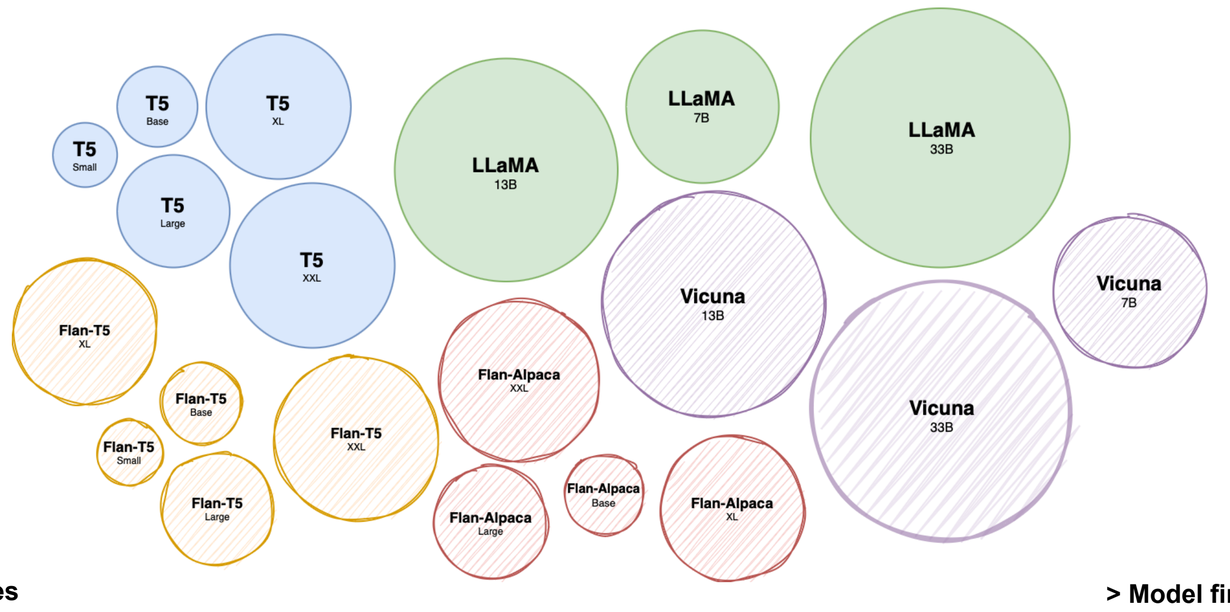
1 Pretrained Language Models

Input:
The use of renewable energy is growing

Language Model

Output:
rapidly in relevance and importance as the world looks towards …

> Model trained to continue sequences



2 Instruction-Tuned Language Models

Instruction:
Write a paragraph about the given topic.
Input:
The importance of using renewable energy.

Instruction-tuned Language Model

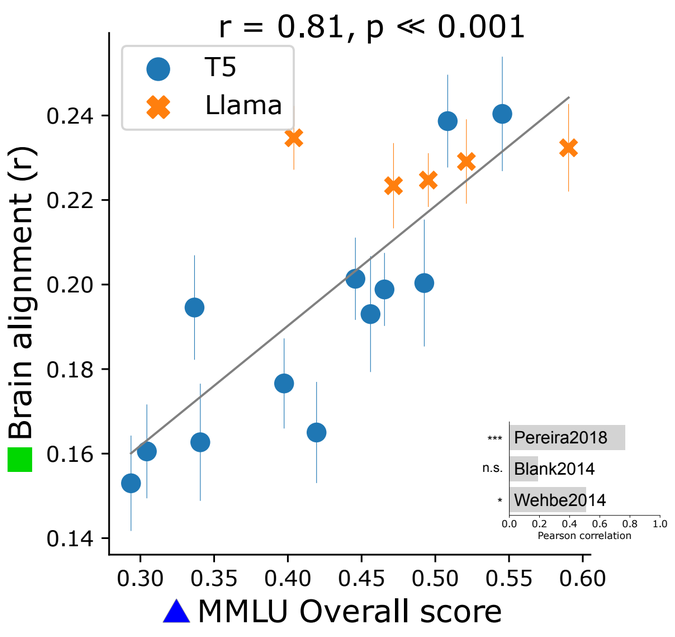
Output:
The use of renewable energy is growing rapidly in relevance and importance..

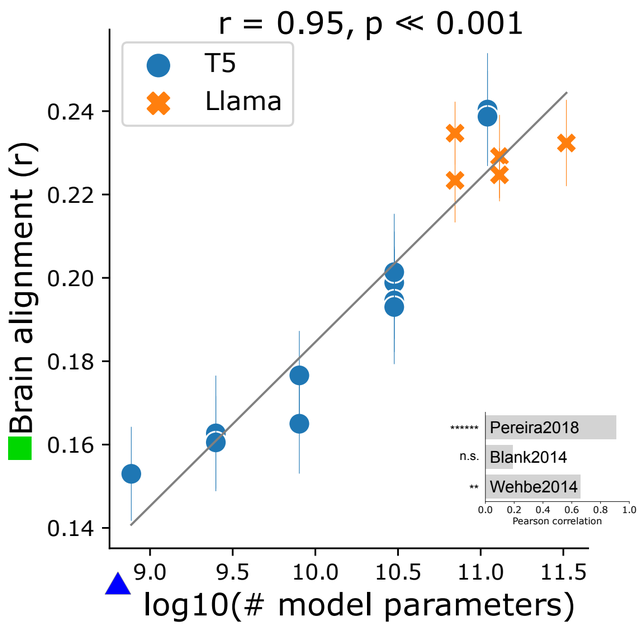
> Model fine-tuned to provide responses to queries

Results

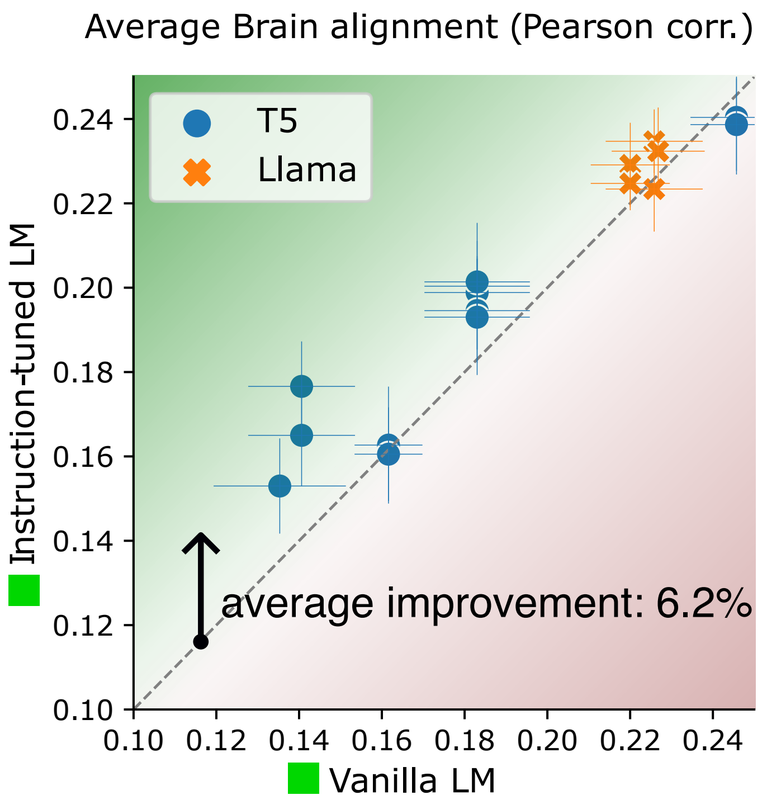
■ Brain alignment ▲ LM properties

Brain alignment correlates with MMLU & Model Size

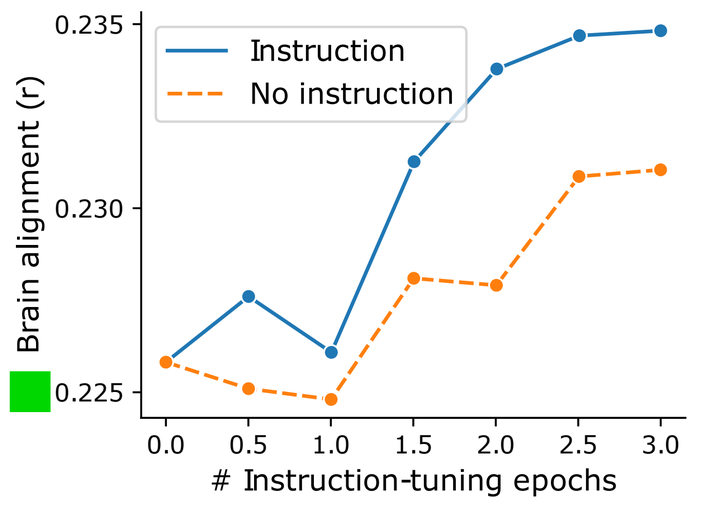


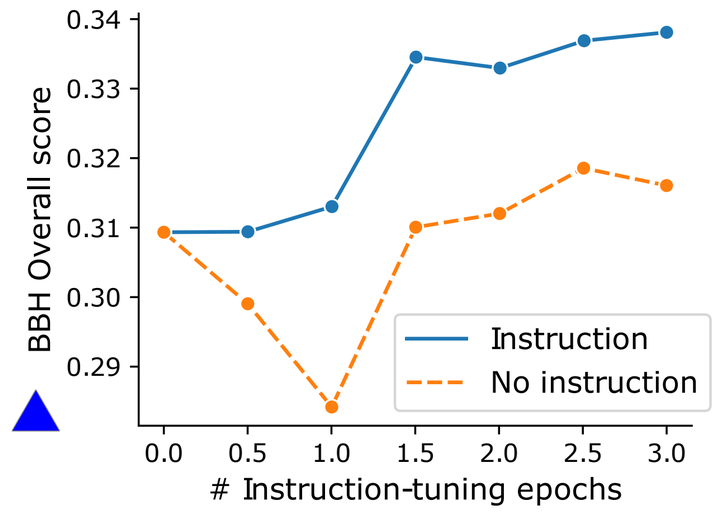


Instruction tuning improves model brain alignment by 6.2% (avg.)



Improvements are due to both training data and process of instruction-tuning





Correlation between brain alignment and performance **MUCH** higher on **world knowledge** benchmarks

Task category	Brain Alignment Correlation (r)	Corrected p-value	Number of tasks	Average (↑) Performance
MMLU – Overall Score	**0.809**	**0.000329**	57	0.36
BBH – Overall score	0.384	0.177	23	0.28
BBH – Algorithmic reasoning	0.194	0.558	8	0.22
BBH – Language understanding	0.163	0.585	3	0.43
BBH – World knowledge	**0.679**	**0.005**	5	0.36
BBH – Multilingual reasoning	-0.035	0.895	1	0.19
BBH – Others	0.478	0.083	6	0.27

← Paper Link

@Khai_Loong_Aw

@bkhmsi